ROADMAP TO

AN INFORMATION & DATA CENTER

Bart Goossens and Dimitri Brosens

Research Institute for Nature and Forest
Klipiekstraat 25
Brussels, Belgium

Abstract: The need for open information recourses becomes inevitable in a modern scientific world. The INBO, Research Institute for Nature and Forest, a scientific institute of the Flemish government decided to develop a Data Resource and Information Centre for all terrestrial biodiversity data in the Flemish region. The development of a modern data centre requires an integration of many different and unforeseen tools. The Data centre story faces integration of the library, assessment and finding old forgotten datasets, data policy implementation, metadata & data indexing and acquisition and co-operative effort between most of the Flemish Institutions and universities. The final goal will be to connect the data centre with the Global Biodiversity Information Facility and to create a gateway for communicating and sharing biodiversity data on the Internet.

Keywords: Information; Biodiversity; Data; Indexing; Metadata; Flanders.

Introduction

The need for open information resources becomes inevitable in the changing modern scientific world. In response, INBO decided to establish a Data Resource and Information Centre for all terrestrial and freshwater biodiversity data in the Flemish region. This roadmap will try to give you a better insight into the development of an information & data center (IDC). We will start with a short introduction of our research institute, then give you a broader view into the establishment of the information & data center.

Introduction to INBO

The Research Institute for Nature and Forest (INBO) is a scientific institute of the Flemish Government. It is a research and knowledge center for nature and forest and its sustainable management and use. INBO conducts research and supplies knowledge to all those who prepare or make policies in this area, and provides biodiversity data to the public.
INBO has branches in Brussels; Geraardsbergen; Grimminge; Groenendaal; and Linkebeek. The main site is located in the heart of Brussels close to Brussels South

railway station. The other sites are spread all over Flanders and are technical satellites with laboratories, tree nurseries, fish farms, etc.

Organization of INBO:
INBO is a very young institute and was established in 2006 by the merger of the Institute of Nature Conservation (IN) and the Institute for Forestry and Game Management IBW). On 1 November 2007 a new administrator-general was appointed, Dr. Jurgen Tack. His first task was to implement a new organizational structure, which became operational on 1 January 2009. Due to this merger INBO has become the leading scientific biodiversity institute in Flanders end employs 250 staff, mainly researchers, information technologists and technicians.

INBO recently identified twelve strategic goals. These goals can be considered general rules and best practices that should be implemented and fully applied by 2015. They can be separated into 6 institutional and 6 scientific goals. Three of the latter concern biodiversity and the natural environment, and the remaining three deal with research on the management and sustainable use of biodiversity.

Due to this central position in the information stream, the IDC is related to each of the twelve strategic goals, but for now we'll focus on the six institutional goals because they have a big impact on the activities of the IDC.

We want to be a high-performance institution that coordinates nature and forest research to reach the whole Flemish community (from citizens to farmers to policymakers). Providing advice to scientists and policymakers influences the actual policy on nature and forestry in Flanders. Besides that we report on the current situation of nature and forest, and of course dissemination of the results of the research activities is an important goal.

The sixth institutional goal, "INBO manages and discloses [biodiversity] data," applies entirely to the IDC. For this, five operational goals were defined. These goals can be seen as a practical translation of the sixth institutional goal, and are expressed as specific measurable activities. A further step is linking the job description of every INBO employee to the aforementioned operational goals.

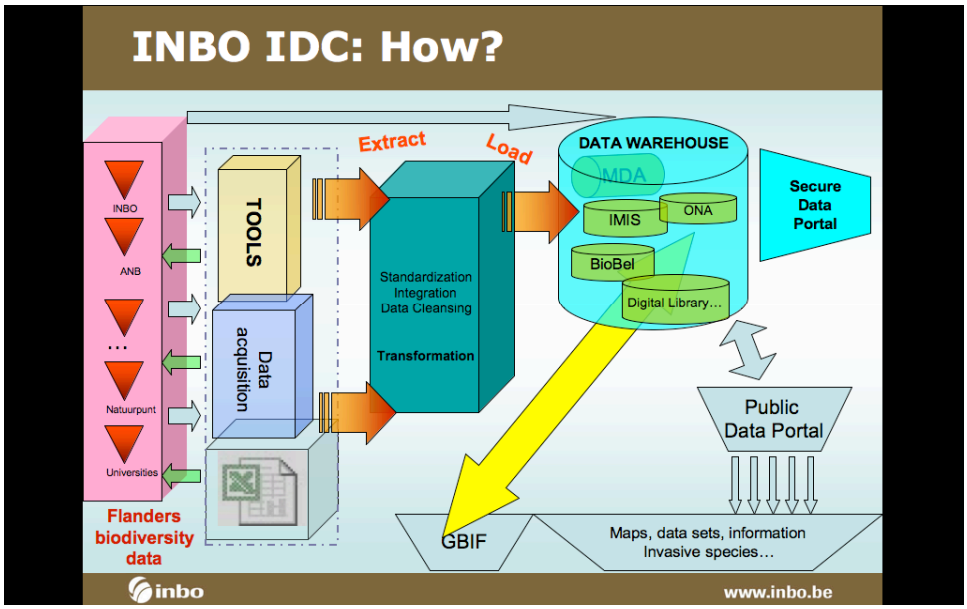
Quality control is one of these operational goals. In this respect, our aim is to distribute highly qualitative and reliable data to the scientific world. In order to do this we have to centralize, preserve, document, integrate and disseminate data in an integrated information system. Data acquisition protocols have to be adapted and implemented. IDC will coordinate the acquisition and storage of data.

Centralization and archiving of reliable data
A clear distinction between data, metadata and physical and digital documents must be made. Data (measurements, observations) need to be structured and stored in databases. Metadata related to scientific datasets and metadata on persons, institutes, projects, events, etc will be integrated into one central information system.

Documentation & structuring of available data
Of course the available data must be documented and structured. Integration of distributed data in the framework of global analyses
will not only give INBO the opportunity to match separate data, but will also provide easy access to different types of data.

Disclosure of data to INBO users and beyond
Every human being (especially those residing in Flanders) is a potential user of INBO data and information. As such, INBO does not only intend to reach out to researchers of INBO and other scientific institutes, but also to policy makers (our bosses), and the general public (who provides us indirectly with the financial means via taxes). And the information INBO wants to provide is very diverse, from invasive species to species protection, habitat restoration to pollution (like the problem with eel pollution).

Structure of INBO
INBO is divided into three sections. There are two research departments divided into 'Biodiversity and Natural Environment' and 'Management and Sustainable Use,' each with five research groups and a 'general department.' The general department offers services first of all to both research groups and to the public.



The IDC is structured into three teams with specific roles. Of course the IDC will only be effectively operational when there is good interaction and cooperation between these teams.

Team Data will set up a data policy that will be used as a standard for the acquisition and management of the data. An empty data center is worth nothing, so numerous contacts and relations have to be made with researchers in Flanders. We will have to convince scientists of the added value of having their data integrated in the IDC. Another high priority is data quality and validation, and this will also be under supervision of this team.

Team Information has to focus on information management. Basic tasks include planning of acquisitions (which materials the library should acquire, by purchase or otherwise), library classification of acquired materials, preservation of materials (especially rare and fragile archival materials such as manuscripts), the deaccessioning of materials, patron borrowing of materials. These basic aspects of library management must be expanded with new tasks like the integration and dissemination of different types of information.

There will be numerous digitization projects and dissemination of our own research information via repositories will be promoted. New library services and web services have to be explored and implemented. Team IT has to maintain these information systems and they will develop the new tools required by the other teams.

The establishment of a collaborative management structure to coordinate and guide the implementation and ongoing maintenance of the information center must include:

- Setting policy regarding participation, funding, development and access;
- The adoption of common standards and best practices to ensure full informational capture;
- Guaranteeing universal accessibility and interchangeability;
- Simplification of retrieval and navigation;
- Facilitating archivability and enduring access.

Different technologies are required: open linked data (interlinking data with other resources), open url generator, Service Oriented Architecture and web services, WEB 3.0, and more.

Web Services

Web services are automated information services that are conducted over the Internet, using standardized technologies and formats/protocols that simplify the exchange and integration of large amounts of data over the Internet. They make it easier to conduct work across organizations regardless of the types of operating systems, hardware/software, programming languages, and databases that are being used.

Establishing a data and information center allows us to integrate the digital library and the data section. Integration is very important because we need to integrate as many metadata and different types of data (such as projects, conferences, datasets, etc.) as possible into one integrated information system. With the IDC we want to create a portal that will be kind of a virtual shopping mall where scientists can find:

- Raw data ((Species, Place, Date, collector)
- Data about raw data (metadata)

- Related information (links between persons, publications, institutions, data sets, events etc.)
- The VLIZ IMIS-system (Integrated Modular/Marine Information system)
- Biodiversity maps with different layers (cf. Geographic Information System) and options (more species, habitats, natural reserves, Ramsar areas etc.) able to generate intuitively usable information.

The data will be an important source for scientific research: scientists can download biodiversity data directly from the IDC and will be able to use the data in their scientific work. So, there will be a reduction in cost for fieldwork (as an example), and more availability of historical data.

In the IDC we can visualize and quantify biodiversity research in Flanders, and also comply with some goals of international conventions like CBD (convention on biological diversity) and European directives.

Next Steps
We just discussed the outlines of the upcoming information and knowledge center. The next question we need to solve is: how are we going to build it? Since we don't want to reinvent hot water, we studied existing biodiversity information initiatives and started a partnership with VLIZ, a famous marine data center in Flanders. First we asked ourselves some questions that are the baseline of the knowledge and information center.

To begin, in the pre-project phase we defined what the needs were: creation of an information and knowledge center where different stakeholders could find lots of different data, from very raw data to more specialized and enhanced data; from data with a direct link to policy to data that can be reused in basic biodiversity research. We also tried to find out what is possible in data warehousing related to biodiversity and science. Further, we agreed on what we should buy, and what we can develop ourselves.
As our partner, VLIZ will help us to construct the very basics of the knowledge center and together we have created a timeline. So there are concrete plans in place.

Different roles will be given to different partners. Privileged partners will have more options then other visitors and will be able to extract more data then unregistered members. For instance, a passing visitor can find a lot of biodiversity information, from diversity maps to species list to policy related information. A registered visitor will have the option to download certain biodiversity datasets, enhanced with a certain amount of metadata. Because some of the data is not freely accessible and is owned by other agencies, an option will be created where the policy of the other agencies will be followed. This will be the case for data in the data center that is not freely accessible and must be purchased. Here INBO is primarily the host for the data.

The data warehouse will be the very heart of the Information and knowledge center. We will integrate library systems (persons, publications, library services, events, etc.) (IMIS system developed by VLIZ) with archiving structures (ONA & MDA) & data systems

with defined structure and web services. The information & data center will have an important role in archiving and distributing datasets.

SLIDE

We will get the information where it is. We estimate that most of the Belgian biodiversity data is in one way or another related to INBO. Then the data will be standardized by the use of a number of IT tools. An example of such a tool is Recorder 6. This tool can be used to clean, validate and standardize biodiversity datasets.

Once all these processes are finished, all the data will be loaded into the data warehouse. A copy of the original data will be kept in the archive, so the original data will always be traceable. Once there is data in the warehouse it can be extracted and used for different options. These include:

- A gateway to GBIF, the global biodiversity information facility.
- A public portal, where a certain number of datasets are available for download.
- Some options related to the generation of maps and policy related issues.
- In the secured data portal even more options are possible.

First Steps

First we need to agree on different standards. Based on these standards we will develop and implement an internal data policy (INBO data only). New procedures are in development and review for creating and submitting data sets and publications. At a later stage we will implement the external data policy, which will be slightly different and will depend on partners and agreements. Another baseline is the Belgian (Flanders) Species list which will be finished soon. Without a species list, there can be no validation or quality control.
Meantime a survey and description of INBO data have been started and VLIZ and INBO are working on adapting and synchronizing information systems. The validation of datasets can only start when the species list is approved. (Recorder 6 proposed)

Filling up the Warehouse

INBO already uses a version of the IMIS (Integrated Marine Information System) system in the Library. For the IDC a new version of IMIS will be developed in cooperation with VLIZ and the data transfer of the INBO library (publications, persons, institutions, projects and events) should be relatively easy. The species list and the data set will then follow.

Meantime we hope to collect all much of the biodiversity data in Flanders. Before filling the warehouse we absolutely need a confirmed species list, then the real data collection and validation can start. Other datasets will also be collected in Flanders. Possible sources are of course the INBO, but also universities, other institutions, workgroups etc. Problems will appear during the collection of the data. Belgium is a federal country and natural conservation is under control of the regions. This means that the only biodiversity data INBO has are data originated from Flanders. Since species are not interested in

regional boundaries, and many datasets are originally from Belgium, collaboration between the regions needs to be established.

Although we will begin with INBO because we know that 80% of the biodiversity data in Flanders is in one way or another related to INBO, we will later visit other institutions to promote the advantages of data sharing in the IDC. Of course, provided data will always be owned by the data creator.

If we look at INBO, and this will probably be the same for many other institutions, we can divide the data in three subcategories. The first category we name Death Data. These can be described as all the unused datasets on a system or network. This is mostly small project data that is not integrated in a big dataset or that may even be lost in the system. Historically, the metadata are mostly incomplete since their importance was underestimated in the past.

One time consuming method of finding Death Data is of course backtracking data through old publications and persons. In INBO we tried to find an automated tool capable of finding unknown datasets on the system (SERVERS and shared computers). We contacted several potential partners. Some solutions could be solved by the Google Search Application enhanced with VLC. Another promising tool was Trillium but it also was limited. The perfect tool has not yet been discovered and probably does not exist at the moment.

The second category is Dormant Data – known data that are not yet digitized, such as collections and literature. Most of the INBO data are digitized. Digitizing data, mostly in other institutions, is a task for the Belgian Biodiversity Platform. This is a federal science policy interface facilitating biodiversity research in Belgium. Some ongoing projects are the Flora databank, the Dipol 2 projects and the Beetle project.

The third category is Living Data - regularly updated datasets or databases currently in use for scientific purposes. The hunt for living data is our main goal. We do this through data archaeology, digging in the past of the institution. Starting from interesting publications, we try to contact the authors in the hope to find relevant datasets. In this way we hope to complete the INBO metadata that can then be incorporated into the system. This is very time consuming, but it is the most effective way.

For describing data there are still some options open. A new tool just developed by GBIF is excellent for describing biodiversity metadata. The tool is free available on http://www.GBIF.org. It will probably become the standard for describing biodiversity metadata. In IMIS there is also a module for describing datasets that can be used for biodiversity data, but also for other type of datasets.

The Future

The IDC will hopefully improve research quality in Flanders. The plan is to study biodiversity from a holistic viewpoint. Many different data types can be compared

without too much difficulty. The data will be centralized and safe, and will be automatically updated to new database standards.
Some research will become a lot cheaper and it will be possible to achieve results more quickly. For some questions, no extra fieldwork will be needed because the data are already there. Easy options include comparing old and new data; comparing the same species from different regions; simple extract of migratory patterns from different datasets; and easy discovery of species interactions. The options are endless!

Vertel hier maar wat over de voordelen!